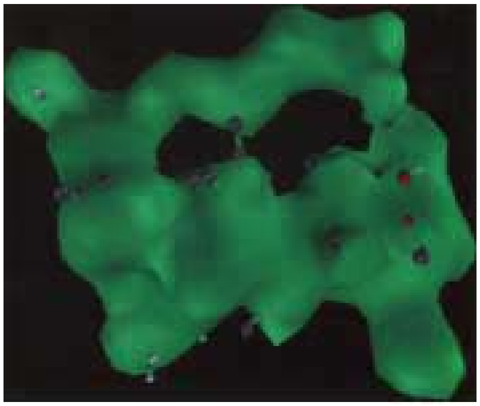
Homology Modeling in the Bioinformatics Age

A Research Project of Dr. Raul E. Cachau

Proteins are essential chemical components of every known form of life. Their wide functional ability is not the result of chemical complexity but, instead, that of nature's ingenuity to build complex structures from simple components. Proteins are linear polymers, based on merely 20 different building blocks that define the sequence of the protein chain. However, if the three dimensional (3D) structure of a protein would resemble that of a loose rod, few different functions could be expected from it. Protein functional flexibility results from the ability of the chain to fold over itself, defining complex 3D objects.



F-1. The X-ray restraint consists of a low resolution electron density map. In a high resolution map the structure fits the features of the density at atomic resolution. In a low resolution map only the seconday structure elements, preserved among homologous molecules, are clearly distinguishable.

Structural studies of biomolecules have changed our perception of the biological world in the past twenty years. Today the explanation of a biological function seems incomplete unless the shape of the biomolecules involved is included in the description. In spite of its relevance and importance, biomolecular structure modeling remains one of the most complex and challenging tasks in molecular biology. One technique used to meet this challenge is homology modeling.

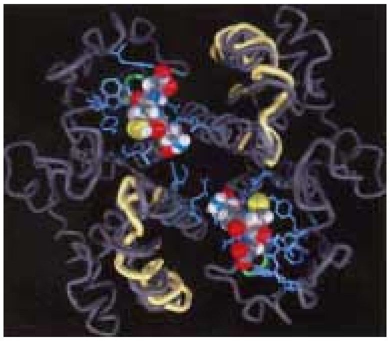
Homology modeling describes an extended collection of techniques with the goal of predicting the 3D details of biomolecules of unknown structure, relying heavily on resources such as pattern-to-function relationship predictors and sequence-to-structure determination predictions. Thus, homology modeling techniques are directly impacted by the current explosion in sequence and structure databases.

The studies carried out in the past ten years have resulted in a wealth of information both at the sequence and structural levels. As the Protein Data Bank (PDB) readily approaches 10,000 structures, it has become evident that the number of 3D arrangements used in naturally occurring proteins is rather limited. Even though many new structures are being added to the databank, the number of different structural motifs in the database is not increasing at the same rate. This limitation of structural motifs means that each motif is being repeatedly sampled in the structural databases. The use of more reliable sequence alignment and 3D-to-sequence alignment techniques is opening the door to the possible prediction of partial structural motifs with a high degree of certainty. The combination of bioinformatics techniques with state-of-the-art modeling tools is potentially one of the most exciting developments in recent years in the area of biomolecular structure studies.

The studies carried out in the past ten years have resulted in a wealth of information both at the sequence and structural levels. As the Protein Data Bank (PDB) readily approaches 10,000 structures, it has become evident that the number of 3D arrangements used in naturally occurring proteins is rather limited. Even though many new structures are being added to the databank, the number of different structural motifs in the database is not increasing at the same rate. This limitation of structural motifs means that each motif is being repeatedly sampled in the structural databases. The use of more reliable sequence alignment and 3D-to-sequence alignment techniques is opening the door to the possible prediction of partial structural motifs with a high degree of certainty. The combination of bioinformatics techniques with state-of-the-art modeling tools is potentially one of the most exciting developments in recent years in the area of biomolecular structure studies.

Some three dimensional information is necessarily lost in the process of 3D-to-sequence alignment, however. This loss of information frequently results in errors in the final model. To overcome some of these difficulties we have developed a series of techniques that map 3D to 3D information directly. One of these techniques is based on the molecular replacement strategy frequently used in X-ray crystallography to overcome the crystallography phasing problem. The 3D coordinates of a molecule of similar composition to that structure which is being solved is used as a probe to find the placement of the new structure in the crystal unit cell. From this initial borrowed model, a new structure is refined. In a novel approach to the homology modeling problem we have shown that this crystallographic technique can be reverted; starting from a known structure, the 3D model of a protein of known sequence and unknown structure can be built. The technique is based on the same assumption used in X-ray crystallography -- that at low enough resolution two homologous structures have a similar 3D structure. This technique takes advantage of the fastest growing segment of the Protein Data Bank. This subset of the Protein Data Bank contains the experimental observables used for the generation of the X-ray crystallography models. Ten years ago less than one in every ten structures in the PDB were accompanied by the experimental observables. Today, almost two in three structures deposited in the PDB contain this information. Use of the crystallography observables may offer a new venue to more reliable homology modeling strategies.

In addition to an overall shape restraint, a biased molecular dynamics trajectory developed in our laboratory is used to add detailed information to the model. During the trajectory, our procedures introduces a subtle bias which forces the model to a conformation similar to that of the parent molecule. This is accomplished by using automatically scaled restraints during the MD trajectory. The scale factors are determined as a function of the agreement of the modeled structure with the parent one, thus preserving as much of the initial model as possible.



F-2. The figure depicts the model of Glutathione S-Transferase (GST) of the Mu class with the parent (gray) and the model structures in yellow and green. Only the regions with larger differences are displayed. The main deviations are located around the xenobiotic receptor area. The region binding glutathione (CPK model) shows very small differences across families of the Mu class.

These techniques help us identify structure-activity relationships in large complex systems like the homodimers of human Glutathione transferases of the Mu class (Figure F-2), Placitaxel, heat shock proteins, and a number of other targets including membrane proteins and aspartic proteases. These methods are also being applied to the initial stages of protein structure refinement by X-ray crystallography techniques.

Benefits of Scalable Increases in Compute Power

Although highly computationally intensive by current standards, the techniques developed in our laboratory have proven valuable in increasing the accuracy of the final models. As our understanding of simpler biological systems grows, it becomes evident that more accurate modeling protocols are needed before we can confront the intrinsic difficulties of highly complex systems. In order to reach this goal a large increase in computational power is needed. High throughput CPUs and large memory architectures are required to pursue complex self-consistent 3D modeling strategies. A two order of magnitude increase in computer performance will enable near optimal multiple sequence alignments to detect long distance homologies, identify hypervariable and conserved regions. This will enhance our ability to identify the initial secondary structure of proteins and build new models using the available 3D information from known structural databases.